DOCUMENT RESUME

ED 364 593 TM 020 837

AUTHOR Galarza-Hernandez, Aitza
TITLE What Is the Probability of Rejecting the Null Hypothesis?: Statistical Power in Research.
PUB DATE Nov 93
NOTE 30p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (22nd, New Orleans, LA, November 9-12, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Behavioral Science Research; Editors; *Estimation (Mathematics); Hypothesis Testing; Literature Reviews; *Probability; Research Design; *Research Methodology; Scholarly Journals; *Statistical Significance
IDENTIFIERS *Null Hypothesis; *Power (Statistics)

ABSTRACT

Power refers to the probability that a statistical test will yield statistically significant results. In spite of the close relationship between power and statistical significance, there is a consistent overemphasis in the literature on statistical significance. This paper discusses statistical significance and its limitations and also includes a discussion of statistical power in the behavioral sciences. Finally, some recommendations to increase power are provided, focusing on the necessity of paying more attention to power issues. Changing editorial policies and practices so that editors ask authors to estimate the power of their tests is a useful way to improve the situation. Planning research to consider power is another way to ensure that the question of the probability of rejecting the null hypothesis is answered correctly. Four tables and two figures illustrate the discussion. (Contains 28 references.) (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 364 593

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
AITZA GALARZA-HERNANDEZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# What is the Probability of Rejecting the Null Hypothesis?: Statistical Power in Research

Aitza Galarza-Hernández

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Mid-South Educational Research Association, New Orleans, November 12, 1993.

TM020837

Abstract

Power refers to the probability that a statistical test will yield statistically significant results. In spite of the close relationship between power and statistical significance, there is a consistent overemphasis in the literature of statistical significance. This paper discusses statistical significance and its limitations and also includes a discussion of statistical power in the behavioral sciences. Finally, some recommendations to increase power will also be provided.

Cohen's (1962) study has become a classic in the area of power analysis. After analyzing published studies in the area of abnormal and social psychology to determine the statistical power of its tests, Cohen (1962) found that the studies in question "had, on the average, a relatively (or even absolutely) poor chance of rejecting their major null hypothesis..." (p. 151). Even though some attention has been given to statistical power issues in research after the publication of Cohen's classical study, Sedlmeier and Gigerenzer (1989) found that 24 four years after its publication no increase in the power of tests have been reported on studies in the field of abnormal psychology. It is evident that in spite of the importance of power analysis, researchers seem to neglect it when conducting research. This is contrasted by an overemphasis of statistical significance issues (Chow, 1988; Cohen, 1962, 1990; ). However, power is intimately related to statistically significance. In fact, power can be defined as the probability of obtaining a significant result (Cohen, 1992). The purpose of this paper is to discuss statistical significance and some of its limitations as well as to highlight the importance of power analysis in the behavioral science research. A secondary purpose of this study is to briefly explain the relationship between power analysis and statistical significance. The power of two studies published in the Journal of School Psychology will also be presented along with recommendations on how to increase the power of statistical analysis.

Some researchers have attempted to explain the overemphasis of statistical significance while power analysis is, basically, neglected. For instance, Cohen (1962) explained that the neglect of power issues originate in the graduate training

of investigators. Graduate statistical textbooks "characterized by an early introduction to statistical significance and power followed by a neglect of the latter throughout the remainder of the text. Thus, every statistical test is described with careful attention to issues of significance, and typically no attention to power" (p. 145).

In reality, a discussion of power analysis will not be complete without referring to statistical significance and of its limitations and misconceptions. As was noted before, power and statistical significance are closely related. Nevertheless, this relationship does not justify the overemphasis of one over the other.

Statistical Significance: Misinterpretations, misconceptions and limitations

Statistical significance is achieved in the statistical significance testing procedure when the researcher is able to reject the null hypothesis (Ho). However, to reject the null hypothesis the researcher's obtained p value has to be less than a predetermined value, usually set at the .05 level. This .05 value has been described as the "sanctified" or "magic" .05 level (Cohen, 1990). Rosnow and Rosenthal(1989), discussing some of the implications of obtaining statistical significance, stated:

> It may not be an exaggeration to say for many PhD students, for whom the .05 alpha level has acquired almost an ontological mystique, it can mean joy, a doctoral degree, and a tenure track position at a major university if their dissertation p is less than .05. However, if the p is greater than .05, it can mean ruin, despair and their advisor's suddenly thinking of a new control condition that should be run. (p. 1277)

Even though statistical significance plays such a prominent role in statistical analysis, it has been criticized energetically since the early 1960's (Carver 1978; Cohen, 1962, 1977, 1990; Chow, 1988; Rosnow & Rosenthal, 1989; Thompson, 1987, 1989) mainly because its meaning has been "blown out of proportion". One of these criticism refers to the relevance of the Fisherian legacy (statistical significance testing) to the behavioral sciences. In this regard Cohen (1990) stated that

> the fact that Fisher's ideas quickly became the basis for statistical inference in the behavioral sciences is not surprising--they were very attractive. Take for example, the yes-no decision feature. It was quite appropriate to agronomy, which was where Fisher came from. The outcome of an experiment can quite properly be the decision to use this rather than that amount of manure or to plant this or that variety of wheat. But we do not deal in manure, at least not knowingly. Similarly, in other technologies--for example, engineering, quality control or education--research is frequently designed to produce decisions. However, things are not quite so clearly decision-oriented in the development of scientific theories. (p. 1307)

Therefore, some of the features of the statistical significance testing may not be suitable for behavioral science research as it is currently practiced. In fact, Carver (1978) stated that "educational research would be better off if it stopped testing its results for statistical significance" (p. 378).

Several misinterpretations of statistical significance and its "magical" $p < .05$ value have been identified. First, it is often interpreted that the obtained p value is

the "probability that the null hypothesis is true" (Cohen, 1990, p. 1307), even though it is known that the null hypothesis is never true in the population. In a discussion of this issue Harris (1985) claimed "no one, for example, seriously entertains the null hypothesis, since almost any treatment or background variable will have some systematic effect" (p. 2). In other words, the null hypothesis is always false. If the null is always false rejecting it will not provide us with new knowledge or insights about the research results if the null is rejected.

Second, it is incorrectly believed that "the p value indicates the probability that the differences found between groups can be attributed to chance" (Borg & Gall, 1989, p. 352). A third misinterpretation is that the level of significance indicates how likely is that your research hypothesis is correct" (Borg & Gall, 1989, p. 352).

In statistical significance testing what the obtained p value or p calc really means or represents is "the proportion of the time that we can expect to find mean differences or other tested effects as large as or larger than the particular sized difference we get when we are sampling from the same population assumed under the null hypothesis" (Carver, 1978, p. 382).

Another misconception of statistical significance is its interpretation as the probability of obtaining the same results if the experiment is repeated. Carver (1978) refers to this misconception as the "replicability or reliability fantasy" (p. 385). Carver (1978) further explains that "nothing in the logic of statistics allows a statistically significant result to be interpreted as directly reflecting the probability that the result can be replicate" (p. 386).

In addition to its various misconceptions, statistical significance also has some limitations. One of the biggest limitations of statistical significance is that it is influenced considerably by sample size. Thompson (1989) describes statistical significance "as an artifact of sample size" (p. 66). He further explains and strongly suggests that any decision to reject or not to reject the null hypothesis must be interpreted within this context. Along the same lines Popper (1959) stated "that almost all possible statistical samples of large sample size will strongly undermine a given probabilistic hypothesis" (p. 201). In other words, with a big enough sample any null hypothesis is likely to be rejected and achieve statistical significance as a result (Fagley, 1985).

Conversely, even a large mean difference or other effect will not be detected as being statistically significant if the sample size is small. This phenomenon is illustrated in Table 1. This table presents the results of four hypothetical studies

and their associated t-tests. In terms of mean differences Study 1 and Study 4 are the same, however, Study 4 does not achieve significance because it has fewer subjects. This is known as the "sample-size problem" (Chow, 1988) which poses major challenges to the interpretation of "statistically significant" results.

---

Insert Table 1 about here

---

Considering the evidence previously presented the criticisms against statistical significance are understandable. Cohen (1990) conveys the essence of

these criticisms when he criticizes statistical significance testing, and its purpose as well as the sample size questions:

> The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is *always* false in the real world.... If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is false, what's the big deal about rejecting it? (p. 1308)

Therefore, obtaining statistical significance does not provide us with new information or with ways to interpret the results. The only thing it can be concluded after achieving statistical significance is that "the effect is not nil" (Cohen, 1990, p. 1307). Therefore, statistical significance alone should not be used to do any interpretation of the results obtained.

The reason why statistical significance testing has been critized so much is because researchers have attached without support different meanings and interpretations to statistically significant results. To determine whether statistically significant results have practical or meaningful significance the use of magnitude of effect estimates, also known as effect sizes, has been suggested (Snyder & Lawson, in press). Furthermore, Hill (1990) argues that measures of strength of the effect (which usually are not even reported in experimental articles) might provide a better criterion for judging the significance of a study (p. 668). Therefore, the use of magnitude of effect or effect size measures is recommended as a supplement to statistical significance testing.

Cohen (1988) defines effect size as the degree to which the phenomenon is

present in the population or "the degree to which the null hypothesis is false" (p. 9). The probability that a statistical test will lead to the conclusion that the phenomenon exits in the population is what is known as the statistical power of the test (Cohen, 1988). In other words, statistical power is the probability that a study will yield statistically significant results (Cohen, 1988). However, researchers give more attention to statistically significance and its interpretation than to power analysis, which is what allows them to find statistically significant results. Data analyses are incomplete without reference or consideration of power issues.

The Concepts of Power Analysis

Generally, power is defined as the probability that a statistical test "will yield statistically significant results" (Cohen, 1988, p. 1). More commonly, power is described as the probability of rejecting the null hypothesis when it is false (Hinkle, Wiersma & Jurs, 1988). According to Olejnik (1984) power values ranging from .70 to .85 are acceptable.

At this point a brief reference to hypothesis testing is necessary. McNamara (1990-91) notes that in hypothesis testing there are two distinct types of inferential errors [incorrect decisions] which influence statistical power. These inferential errors are Type I and Type II errors.

---

Insert Figure 1 about here

---

Type I error is defined as rejecting the null hypothesis when it should be accepted. A probability value reflecting the possibility of doing a Type I error can

be related to this incorrect decision as follows (McNamara, 1990-91, p. 27):

p [Type I error]= a

p [rejecting a true Ho]= a

This value also reflects the level of significance and is called alpha (a).

A second error is the Type II error which is defined as the probability of accepting the null hypothesis when it should be rejected (Huck, Cormier & Bounds, 1974). As in the Type I error, a probability value, called beta (ß), can be associated to this type of incorrect decision as follows (McNamara, 1990-91, p. 28):

p [Type II error]= ß

p [Not rejecting a false Ho]= ß

Once the value of ß is established the power of the test can be determined as follows:

Power = p[rejecting a false null hypothesis] = 1-ß

Researchers want to minimize both type of errors and maximize the power of their tests. Certainly, minimizing ß increases the power (1-ß) of the test. However, as ß increases the level of significance decreases. Therefore, researchers need to make certain decisions based, in part, on the objectives and goals of their research when minimizing both Type I and Type II errors while maximizing power (Hinckle, Weirsma & Jurs, 1988). Usually, more care is given to Type I error and significance than to Type II error and power issues (Cohen, 1962). Cascio and Zedeck (1983) state that in current practice researchers would rather make the mistake of failing to find a phenomenon than the mistake of "finding" a

phenomenon that is not there (p. 521). That is, some researchers prefer to have a high probability of a Type II error than a high probability for a Type I error. This could be related to the purposes of the research conducted. For example, if the research has as a purpose determine the effectiveness of a particular intervention program, researchers prefer to say the program was not effective and maybe modify it and try it again rather than saying that the program is effective when it is not.

Several factors have a direct influence on power. Cohen (1977) identifies the significance (alpha) level (a), reliability of measurement, sample size, and effect size (ES) as the factors influencing power. Hinkle, Weirsma and Jurs (1988) add to this list the directionality of the alternate hypothesis (Ha) as a factor which also have an impact on the power of the test. Knowing the value for each one of these factors enables the researcher to determine the power of their test.

The significance (alpha) level (a) is one the aspects in statistical testing and power analysis more discussed in the literature because of its relation with Type I error. Cohen (1988) refers to alpha as the "critical region of rejection" for the null hypothesis. He further warned that for power to be defined, the value of alpha must be set in advance. Hinkle, Weisrma and Jurs (1988) stated that there is an inverse relationship between a and ß. When the values of the other factors are held constant, increasing alpha results in a decrease in ß. Given the nature of the relationship between ß and power (1-ß), a decrease in ß results in an increase in power (1-ß). However, it is common practice in research to find very small alpha values ("the smaller the better") because the researchers are more concerned with Type I error. This results in relatively small power and in an increase in the

probability of Type II errors (ß).

Within the context of significance level, the directionality of the alternate hypothesis (Ha) has been described as a factor that also bears on the power of the test (Cohen, 1988; Hinkle, Weirsma & Jurs,1988). Basically, with all other factors held constant, one-tailed tests are more powerful than two-tailed tests. When conducting two-tailed tests the researcher states that a phenomenon exist if parameters A and B differ. However, no direction of the difference is specified, therefore, departures from the null hypothesis constitute evidence against the null hypothesis and in favor of the phenomenon's existence. In other words, if the null hypothesis can be rejected in either direction this means that the critical significance region will be at both tails of the test distribution resulting in a test with less power due to the fact that both tails need to be tested (Cohen, 1988).

A second factor influencing power is the representativeness of sample results and sample size. Cohen (1988) stated that the "reliability (or precision) of a sample value is the closeness with which it can be expected to approximate the relevant population value" (p. 6). He further noted that it is necessarily an estimate because the population value is unknown. Reliability of sample results and sample size is always dependent upon size of the sample. The larger the sample size, other factors held constant, the smaller the error and the more accurate the results will be. This results in a more powerful test of the null hypothesis or more probability to reject the null hypothesis (Cohen, 1988; Hinkle, Weirsma & Jurs, 1988). In other words, there is a direct relationship between sample size and power, increasing sample size results in an increase of power. Table 2 depicts the relationship

between sample size and power.

---

Insert Table 2 about here

---

The third factor influencing power is the effect size (ES). Cohen (1988) conceptualizes effect size as the most important factor in the determination of power. It was noted previously that effect size (ES) is defined as "the degree to which the phenomenon is present in the population" (p. 9) and as a measure for the determination of findings' practical importance. Hinkle et al. (1988) refer to effect size as the "desired difference to be detected" (p. 306). According to Cohen (1988) this value will be zero if the null hypothesis is true and a nonzero value if the null is false. Power is referred to as the probability of the test to detect this difference.

Effect size has also been described as expressing the discrepancy between Ho and Ha (Sedlemeier & Gigerenzer, 1989). Cohen suggests the use of an effect-size index named d "as a standard which may be used in reporting effect sizes across different studies and research designs" (Arvey, Cole, Fisher Hazucha & Hartanto, 1985, p. 495). This d index represents the mean difference between groups in standard deviation units (Cohen, 1988). The relationship between effect size and power is also a direct one. Sedlemeier and Gigerenzer (1989) stated that "everything else held constant, the greater the effect size the greater the power" (p. 309).

As in the case of significance level, researchers must specify the minimal difference they are interested in finding "a priori", i.e., at the planning stage of the

study. Cohen (1988) strongly encouraged researchers to provide their own definition of a reasonable effect size. This value is particular to each study and depends upon the population, the nature of the variables, the instrumentation as well as the procedure, therefore, effect size determination is a very subjective process (Olejnik, 1984). In fact, effect size facilitates a value judgment on the part of the researcher. However, to facilitate interpretation and comparison between studies Cohen (1977) proposed the use of the effect size index "d". Cohen (1977) also suggested definitions for small, medium, and large effect sizes for different statistical analyses by assigning specific values (in d units). For instance, the proposed definitions for small, medium, and large effect sizes are .20, .50, and .80, respectively. For analysis of variance (ANOVA) Cohen (1988) suggests the values of .10, .25, and .40 for small, medium, and large effect sizes. Even though, Cohen (1988) suggested researchers to provide their own definitions for a reasonable effect size, his definitions of effect size are "the most widely accepted guidelines" (Olejnik, 1984, p. 44).

Power, significance level (a), sample size, and effect size (ES) are intimately related to each other. This relationship is such that any of them is a function of the other three (Cohen, 1988). This relationship allows for different types of power analysis. This paper has concentrated in power as a function of significance level (a), effect size (ES), and sample size (n). A second type of power analysis widely used in research is sample size as a function of power, effect size and significance level. The latter will be described briefly.

To summarize, with other factors held constant, increasing the significance

level, the sample size and using a larger effect size will result in a more powerful test. The relationship between significance level, sample size, and effect size with power are depicted in Figure 2.

---

Insert Figure 2 about here

---

Advantages of doing Power Analysis

Having discussed statistical power, what follows is a discussion of some advantages of using statistical power in research. The literature has identified the special usefulness of statistical power in the planning stage of research (McNamara 1990-91; Olejnik, 1984; Thompson, 1987).

Of course, power analysis is useless following the detection of a statistically significant effect, since a Type II error is impossible in these circumstances. In the planning stage statistical power analysis is especially useful in determining the required sample size (Fagley, 1985). This refers to the power analysis in which sample size is a function of power, effect size, and the significance level (Cohen, 1988). This type of power analysis has been described in the literature not only as very useful to determining sample size during the planning stage of the study (Olejnik, 1984), but also facilitates selecting a design sensitive enough to the differences between the groups (Lipsey, 1990). This becomes especially important when considering that sample size can influence the choice of instrument, design, and analysis (Olejnik, 1984). To facilitate this kind of statistical power analysis Cohen (1977) has designed and published a series of tables that enable the research

to calculate the required sample size given a specified significance level, effect size, and power. A modification of one of these tables is represented in Table 3. Cohen (1977) has also designed these types of tables for all the possible power analyses.

---

Insert Table 3 about here

---

Researchers have stressed the importance of paying more attention to power analysis during the planning stage of research (Cascio & Zedeck, 1983; Cohen, 1988; Hill, 1990;). According to Olejnik (1984) unplanned research is an inefficient use of time and resources to conduct a study. Shavelson (1981) suggests that researchers should "take a power trip". He believes that researcher should strive to design the most powerful experiments. Along the same lines, Borg and Gall (1989) note that the best time for researchers to specify and decide on the actual statistical power for their study is in the research design planning stage. Therefore, the "best practice" in research is to spend some time in the planning stage of research so researchers do not have to deal with results based on low power. Research in which planning for power has been exercised would give researchers a better ground to interpret "significant" results. When conducting a study researchers must specify "a priori" the (a) minimal desired effect size, (b) level of significance, and (c) the desired power (Hill, 1990). Only under these conditions researchers can be assured that their results are interpretable. This is also considered a good strategy for minimizing inferential errors (McNamara, 1990-91).

A second use of statistical power is to evaluate the results of previously conducted research. This analysis refers to the determination of power given a specified alpha value (a), sample size (n), and effect size (ES). In other words, this type of analysis can determine the probability that the study would detect effects of a specified level of alpha, given the sample size and design used (Fagley, 1985). Table 4 illustrated this kind of statistical power analysis with two studies published in the Journal of School Psychology.

---

Insert Table 4 about here

---

Increasing Statistical Power in Behavioral Research

In the introduction of this paper it was mentioned that too often researchers are making conclusions which influence our practices on the field of behavioral sciences based on low power studies. The truth of the matter is that only few articles report power analysis in their studies.

Several researchers are concerned with the status of power in behavioral sciences. Not using power in research studies undermines the findings' relevance of the behavioral science research. Some alternatives have been suggested to maximize statistical power (Arvey, Cole, Fisher Hazucha, & Hartanto, 1985; Cascio & Zedeck, 1983; McNamara, 1990-91). First, the necessity to pay more attention to power issues is observed. Researchers should "take a power trip" as Shavelson (1981) has suggested and consider the power of their statistical tests. It

is necessary to make researchers aware of the importance of power in research.

One way to improve this situation is by changing editorial policies and practices (Thompson, 1987). Sedlemeir and Gigernzer (1989) suggest that the status of power analysis will not change until

> the first editor of a major journal writes into his or her editorial policy statement that authors should estimate the power of their tests if they perform significance testing, and in particular if Ho is the research hypothesis. (p. 315)

The literature also has stressed the importance of planning research. It is being recommended that researchers exercise their ability to reflect upon what it is they want to study, what are the implications of their results. Along the same lines Thompson (1989) states "thinking is always a worthwhile endeavor for researchers and can lead to improved practice" (p. 67). Moreover, McNamara (1990-91) suggests that "the best way of guarding against either type of inferential error is to specify all four essential inference decisions (alpha, beta, Ha, and effect size) in the research planning stage" (p. 32). Therefore, it could be concluded that the planning stage is a crucial part of research and will determine interpretability and usefulness of research planning. This is especially relevant when using power analysis to determine sample size given that "low sample size greatly impaired the power to detect true validity" (Arvey et al., 1985,
p. 494). The same principle applies to the determination of effect size.

A third recommendation to maximize power is related to the level of significance or alpha level (a). Olejnik (1984) suggested that

> since effect sizes in the social sciences tend to be small and sample sizes often cannot be increased greatly a reasonable alternative for maintaining statistical power is to accept an increased chance of Type I error. Over replications of the study, true effects would be separated from Type I errors. This goes in total contradiction to the current practice of overemphasizing Type I over Type II error. Even though it is desirable to minimize the probability of a Type I error, it is also important to have a reasonable probability of identifying a meaningful effect. (p. 47)

Given the relationship between alpha (a) and beta (ß), if we increase alpha, beta will decrease resulting in an increase on power which is our ultimate goal.

Neglect of power issues has gone on for too long. Research findings in studies with low power could be misleading. Power analysis in research is what answers the question: "what is the probability of rejecting the null hypothesis?" As a result of this process we will be able to find a given phenomenon in the population. The question is, how powerful are our statistical analysis to find this phenomena? It is time to give power the attention it deserves.

References

Arvey, R. D., Cole, D. A., Fisher-Hazucha, J., Hartanto, F. M. (1985). Statistical power of training evaluation designs. Personnel Psychology, 38, 493-507

Borg, W. R., & Gall, M.D. (1989). Educational Research: An Introduction. New York: Logman.

Carver, R.P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. Personnel Psychology, 3, 517-526.

Chow, S.L. (1988). Significance test or effect size? Psychological Bulletin, 103, 105-110.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. New Jersey: Lawrence Erlbaum, Publishers.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New Jersey: Lawrence Erlbaum, Publishers.

Cohen, J. (1990). Things I have learned (So far). American Psychologist, 45, 1304-1311.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Fagley, N.S. (1985). Applied statistical power analysis and the interpretation of nonsignificant results by research. Journal of Counseling Psychology, 32,

391-396.

Harris, R. J. (1985). A primer of multivariate statistic (2nd ed.). New York: Academic Press.

Hill, O. W. (1990). Rethinking the "significance of the rejected null hypothesis. American Psychologist, 45, 667-668.

Hinkle, D. E., Wiersma, W., & Jurs, S.G. (1988). Applied Statistics for the behavioral sciences (2nd ed.). Boston: Houghton Mifflin Company.

Huck, S. W., Cormier, W. H., & Bounds, W. G. (1974). Reading statistics and research. New York: Harper Collins.

Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental research. Newbury Park: Sage Publications.

Mattison, R.E., Morales, J., & Bauer, M.A. (1991). Elementary and Secondary socially and/or emotionally disturbed girls: Characteristics and identification. Journal of School Psychology, 29, 121-134.

McNamara, J.F. (1990-91). Statistical Power in educational research. National Forum of Applied Educational Research Journal, 3, 23-36.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Popper, K.R. (1959). The logic of scientific discovery. New York: Basic Books.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an

effect on the power of studies? Psychological Bulletin, 105, 309-316.

Shavelson, R.J. (1981). Statistical reasoning for the behavioral sciences. Boston: Allyn and Bacon.

Smith, M.L., Minden, D. , & Lefevbre, A. (1993). Knowledge and attitudes about AIDS and AIDS education in elementary school students and their parents. Journal of School Psychology, 31, 281-292.

Snyder, P., & Lawson, S. (in press). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education.

Thompson, B. (April, 1987). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868).

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 22, 66-68.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Table 1

Relation Between Sample Size and Statistical Significance for Four Hypothetical Studies

| Study | M1 | M2 | M1-M2 | df | *t*-test significant? |
|---|---|---|---|---|---|
| 1 | 5 | 4 | 1 | 20 | Yes |
| 2 | 12 | 2 | 10 | 20 | Yes |
| 3 | 6 | 2 | 4 | 5 | No |
| 4 | 5 | 4 | 1 | 5 | No |

Note. From Chow (1988), p. 106.

M1= mean experimental condition; M2= mean control of condition; M1-M2=difference between M1 and M2.

Table 2

Statistical Power Estimates for Selected Sample Size, with a Predetermined a1=.05, and a fixed ES=.50

| Sample size | Power (1-ß) | Probability of Type II error (ß) |
|---|---|---|
| 20 | .21 | .79 |
| 30 | .32 | .68 |
| 40 | .45 | .55 |
| 50 | .55 | .45 |
| 60 | .64 | .36 |
| 80 | .78 | .22 |
| 100 | .88 | .12 |
| 180 | .99 | .01 |

Note. From Cohen (1988), (p. 28-29).

As could be observed from the table statistical power is a direct consequence of the actual sample size. With a sample size of 50 the power of the test is .55, that is, the test has slightly more than a 50-50 chance of detecting a true relationship between variables. If the sample size is increased to a 100 the power increased from .55 to .88.

Table 3

Statistical Power Analysis to Estimate sample size with a Fixed Effect Size of .50 and a Predetermined Power of .80

| a1 | Effect size | Power | n |
|---|---|---|---|
| .01 | .50 | .80 | 82 |
| .05 | .50 | .80 | 50 |
| .10 | .50 | .80 | 36 |

| a2 | Effect Size | Power | n |
|---|---|---|---|
| .01 | .50 | .80 | 95 |
| .05 | .50 | .80 | 64 |
| .10 | .50 | .80 | 50 |

Note. From Cohen (1988), (p. 54-55).

a1 = one-tailed test. a2 = two-tailed test

Table 4

Statistical Power as a Function of Sample Size (n), Effect Size (ES) and Alpha level (a)

| Study | Type of Analysis | n | ES* | a | Power |
|---|---|---|---|---|---|
| Mattison, Morales & Bauer (1991) | two-tailed t-test | 65 | .50 | .05 | .80 |
| Smith, Minden & Lefevbre (1993) | Chi-Square | 398 | 30 | .05 | .99 |

Note: The studies mentioned on this table are published in the Journal of School Psychology. As can be observed from the table knowing the type of analysis, sample size, effect size and the alpha level values the power of the tests used can be easily determined by consulting Cohen's (1988) tables.

* A medium effect size was assumed when it was not specified in the method section of the article.

| | **H(O) IS TRUE** | **H(O) IS FALSE** |
|---|---|---|
| **DO NOT REJECT H(O)** | Correct Decision<br>(1-a)<br>Level of significance<br>***Case 1*** | Incorrect Decision<br>Type II Error<br>(ß)<br>***Case 2*** |
| **REJECT H(O)** | Incorrect Decision<br>Type I Error<br>(a)<br>***Case 3*** | Correct Decision<br>(1-ß)<br>Power<br>***Case 4*** |

Figure 1. The Decision Problem in Hypothesis Testing

Note: Mc Namara (1990-91), p. 26. Reprinted by permission.

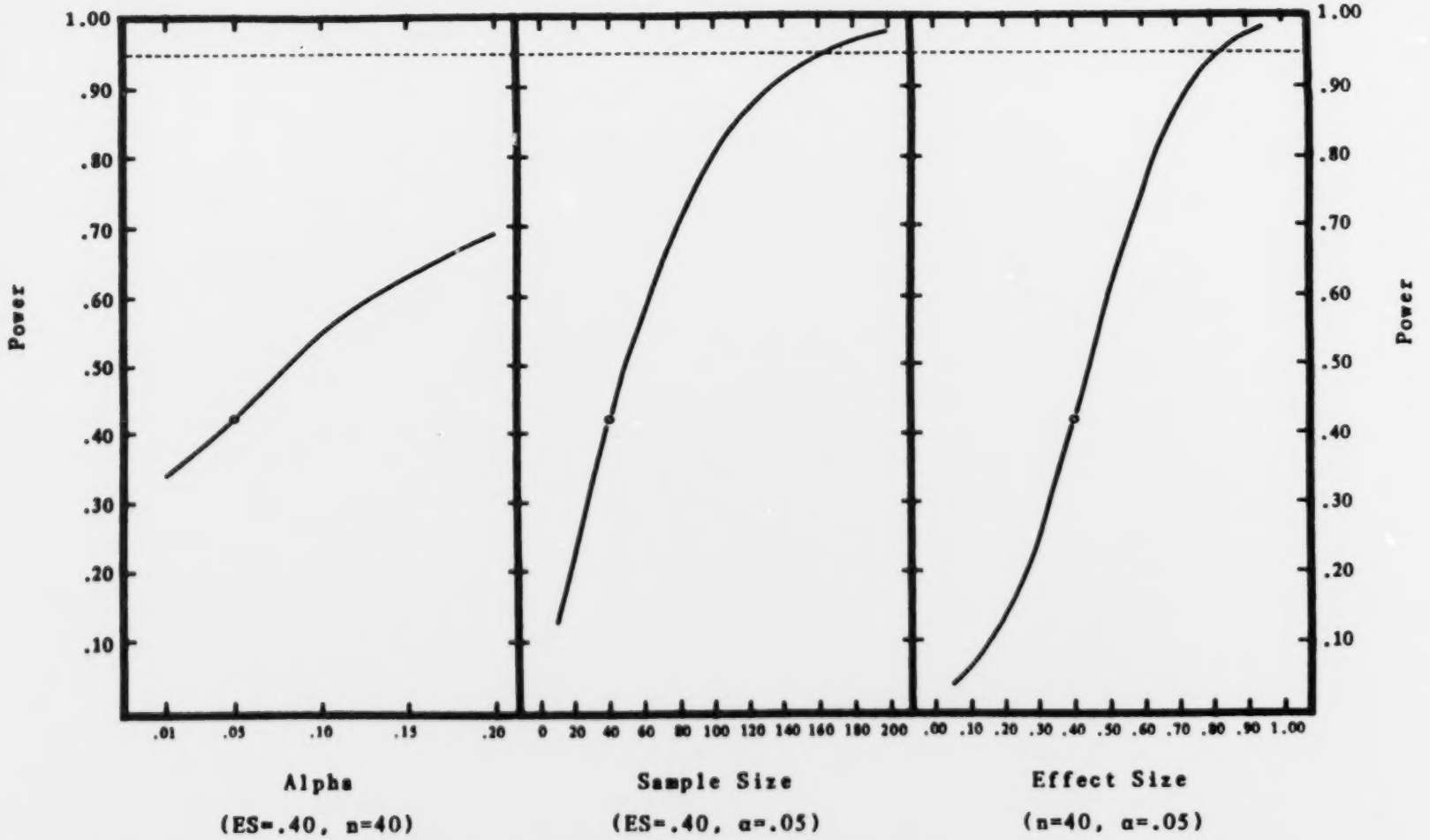


Figure 2. Changes in Power with Changes in a, sample size (n),and Effect Size (ES) for a study (a=.05, n=40, ES=.40)

From Design Sensitivity by M. W. Lipsey, 1990, Newbury Park: Sage Publications.